CAPP 2005
Zeuthen

64-bit Linux

Stephan Wiesand
DESY -DV -

April 7th, 2005

Why would You care ?

- You may have hit the 4 GB limit already
  - if not, you probably will soon
- the fastest commodity (like) systems are now 64-bit
  - best price/performance ratio: dual CPU systems with
    - Opteron
    - Xeon EM64T
    - Itanium II ?
      - not yet, but in 2007 (claim by intel)
      - fast for 64-bit applications only
- this presentation is mainly about AMD64/EM64T
  - also very fast when running existing 32-bit binaries
    - still faster with 64-bit executables

Outline

- brief introduction to AMD64 & EM64T
  - more than just an extended address space
- performance comparisons for physics applications
  - on Opteron, Nocona, Prescott & 32-bit systems
- using linux on these systems
  - 64bit distributions
  - 32bit compatibility
  - Problems
- outlook
  - how to take advantage of the upcoming new systems

AMD64: Terminology

- AMD created this platform under the name x86-64
  - and later renamed it to AMD64
- intel started out with the name IA32E
  - back then, IPF was still called IA64
  - then renamed it to EM64T
    - Extended Memory 64 Technology
- rpm architecture suffix is x86_64 or ia32e
- lately, it's also being called x64 by some
  - java.sun.com, c't, ...
- I'll use **AMD64 as the generic term**

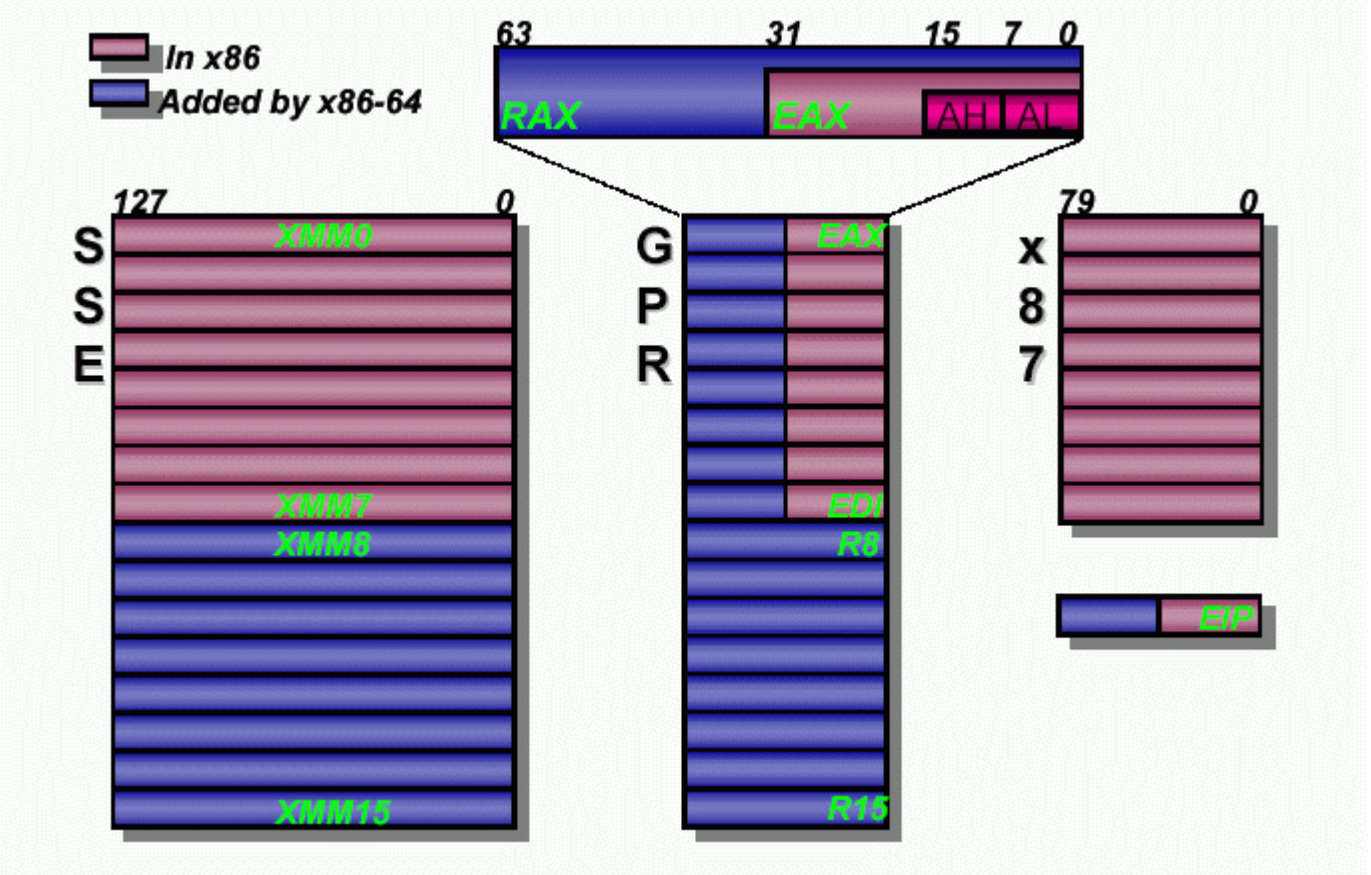
AMD64: Another 64-bit Platform ?

- linux has been running on 64-bit platforms for a while
  - Alpha, Sparc, PPC, PA-RISC, IPF (formerly known as IA64)
  - all are RISC, and none can execute i386 instructions
  - software emulation exists for Alpha and IPF
- AMD64 is an extension of the i386 CISC architecture
  - executes i386 instructions in hardware
  - can run a 32-bit OS
  - supports running 32-bit applications under 64-bit OS
  - 64-bit mode needed an extended instruction set
    - allowed additional registers and addressing modes

64 bits ?

- not quite: AMD64/EM64T support
  - 40 bits (1TB) of physical memory
  - 48 bits (256 TB) of virtual memory
  - current chipsets may support less
    - 915X/925X: 4GB of physical memory...
    - 945G/955X (not yet available): 4/8 GB...
- ABI imposed limits for executables in 64bit mode:
  - "small" code model: 2 GB code + data
  - "medium model": 2 GB code (w/ performance penalty)
- 32bit apps under 64bit OS have full 4GB address space
  - 3GB is the limit under 32bit kernels (3.5 at best)

AMD64 register set



- general purpose registers and instruction pointer are 64 bits wide, twice the number of GPRs
  - all addressable as 8,16,32, or 64 bits as needed
- twice the number of SSE (formerly MMX) registers
  - still 128 bits wide

AMD64 Operating Modes

Operating Mode		OS Required	Application Recompile Required	Defaults		Register Extensions	Typical GPR Width
				Address Size (bits)	Operand Size (bits)		
Long Mode	64-bit Mode	New 64-bit OS	yes	64	32	yes	64
	Compatibility Mode		no	32		no	32
				16	16		16
Legacy Mode	Protected Mode	Legacy 32-bit OS	no	32	32	no	32
				16	16		
	Virtual-8086 Mode			16	16		16
	Real Mode	Legacy 16-bit OS					

- CPU enters Long Mode or Legacy Mode during boot, no way back
- rumour: extended register set could be accessed in 32bit mode as well ("REX32")
  - would still need modified OS and compilers

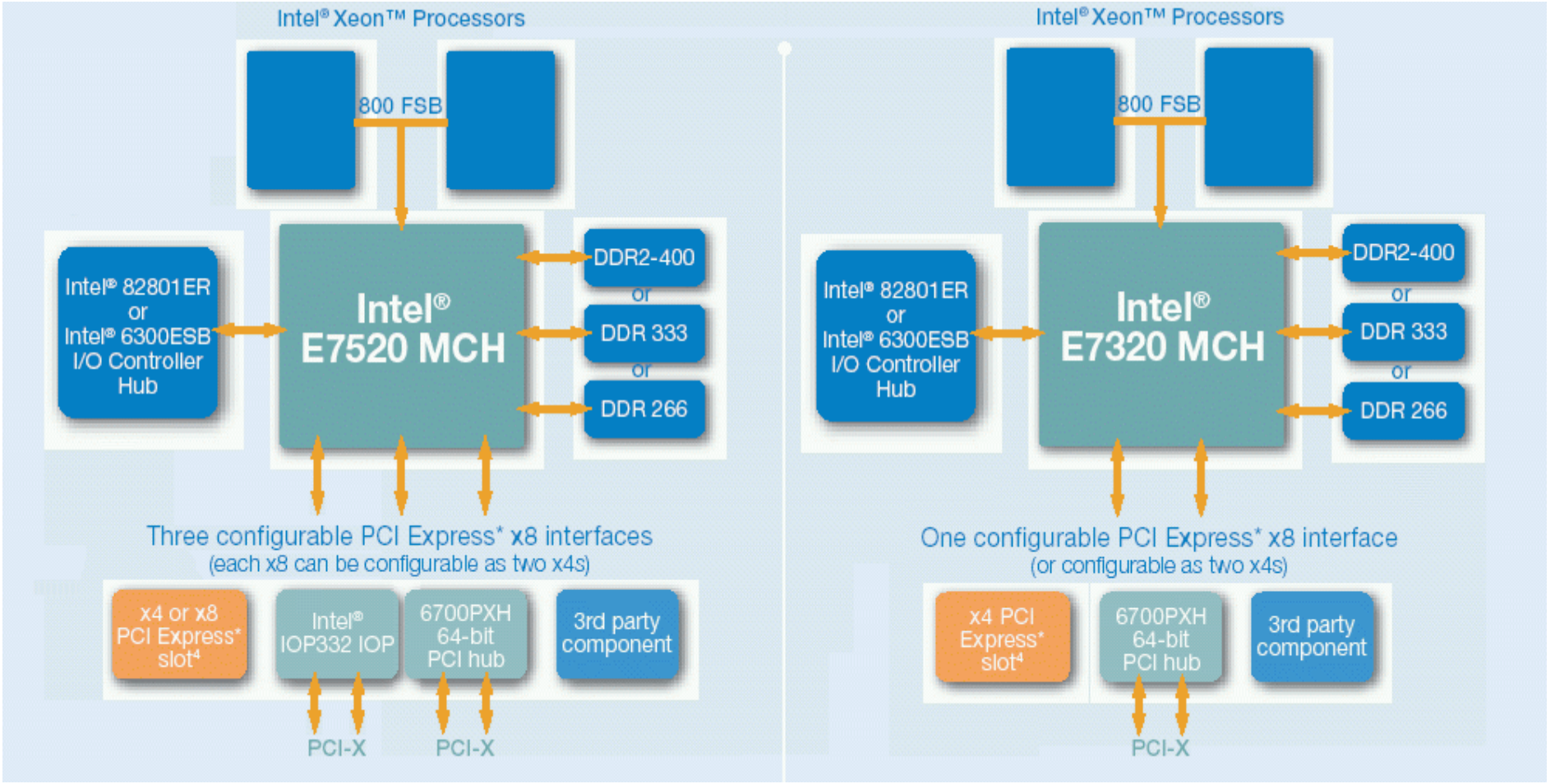
AMD64 Instruction Set Changes

- besides 64bit specifics:
- effective protection of memory against execution
  - "NX" bit
  - available in 32bit mode as well
- generally usable instruction pointer relative addressing
  - reduced performance penalty for position independent code
    - -> shared libs
    - from 20% to 8%
- 64bit apps must not use x87 instructions
  - x87 stack not preserved across context switches

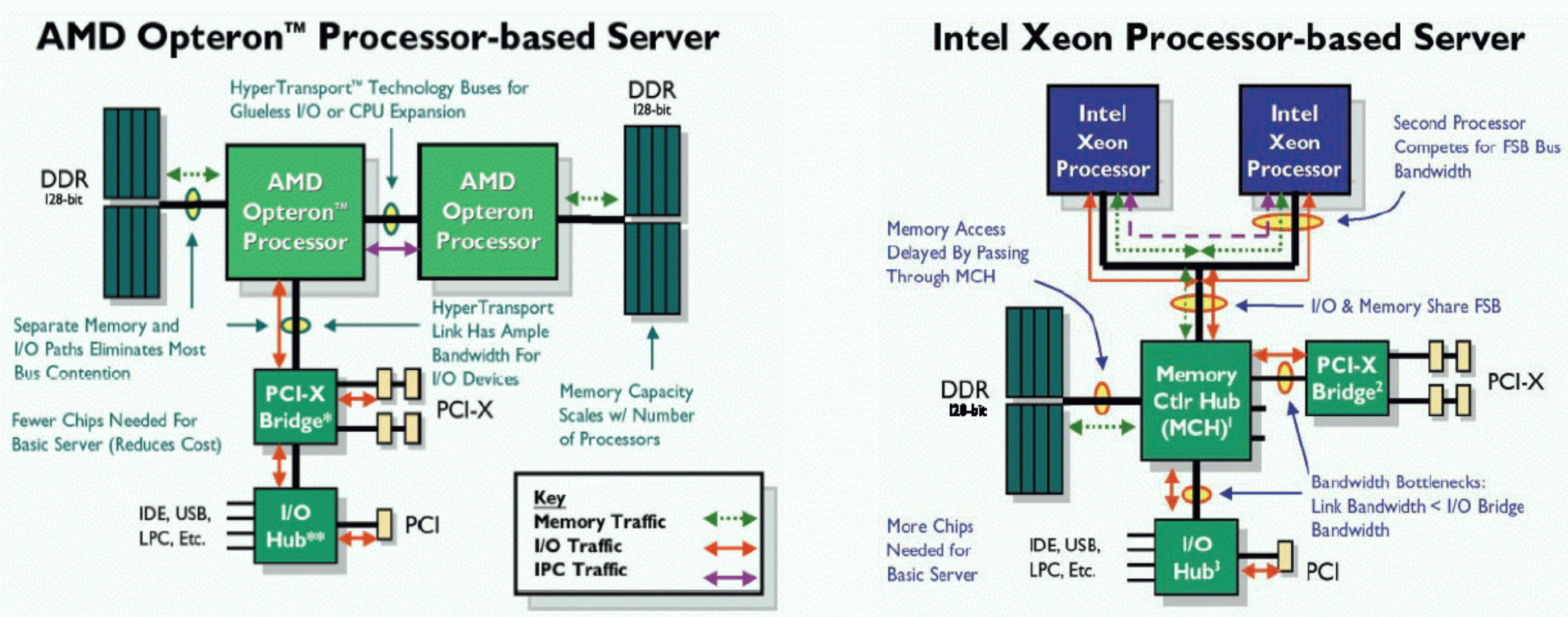
AMD64/EM64T Differences

- most visible: SSE instructions
  - both implement SSE2
  - only AMD64 implements 3dNow!
  - only EM64T implements SSE3
    - yet - 90nm Opterons with SSE3 start shipping now
- a few more subtle differences in instruction sets
  - should only matter for kernel, glibc, compilers
  - should not affect ordinary application programmes
  - everything we compiled with pre-EM64T gcc releases worked on EM64T systems

Where's the Bottleneck ?

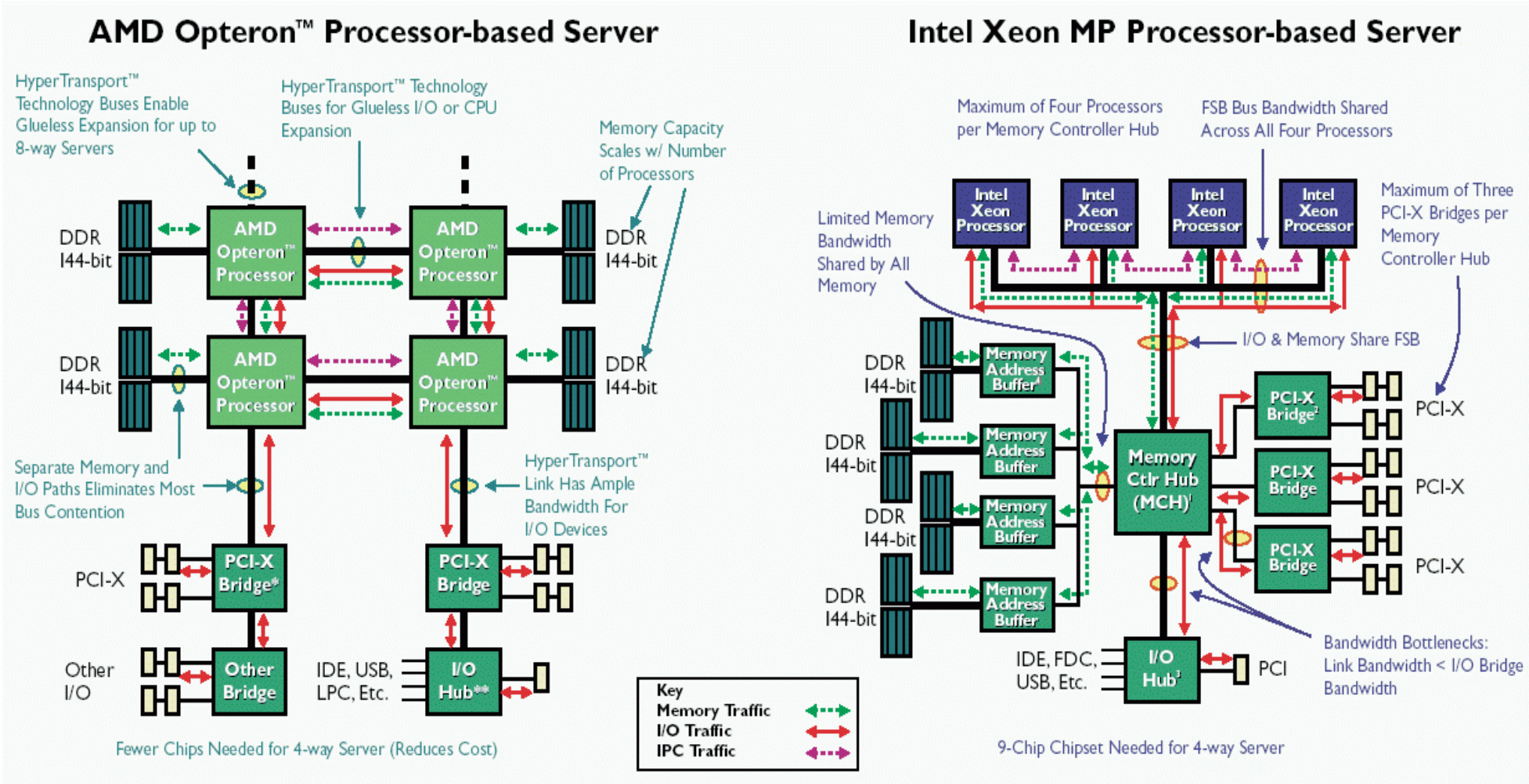


AMD's Additional Step



- currently all 6.4 or 8 GB/s:
  - memory interfaces
  - front side bus
  - HyperTransport links

4-way systems



- we start seeing 4-way systems that make (more) sense
  - and are becoming affordable because there's a sizable market

NUMA: Non Uniform Memory Access

- memory now may be more or less close to CPU
  - cache coherent access to remote memory at full bandwidth
  - but bandwidth has to be shared and latencies increase
  - requires kernel with NUMA support to be most efficient
    - memory should be allocated close to requesting process/thread
    - processes/threads should be scheduled close to their memory
- alternatively, BIOS may also present all RAM to the OS as single uniform block, node memory interleaved by page
  - no OS support required
- whenever using shared memory, allocate it from the process or thread that uses it most

Other Differences in Architectures

- DMA to memory above 4 GB from 32bit I/O devices
  - Opterons have an I/O MMU to make this possible
  - Intel's chips do not
    => have to use bounce buffers

Hardware for Performance Comparisons

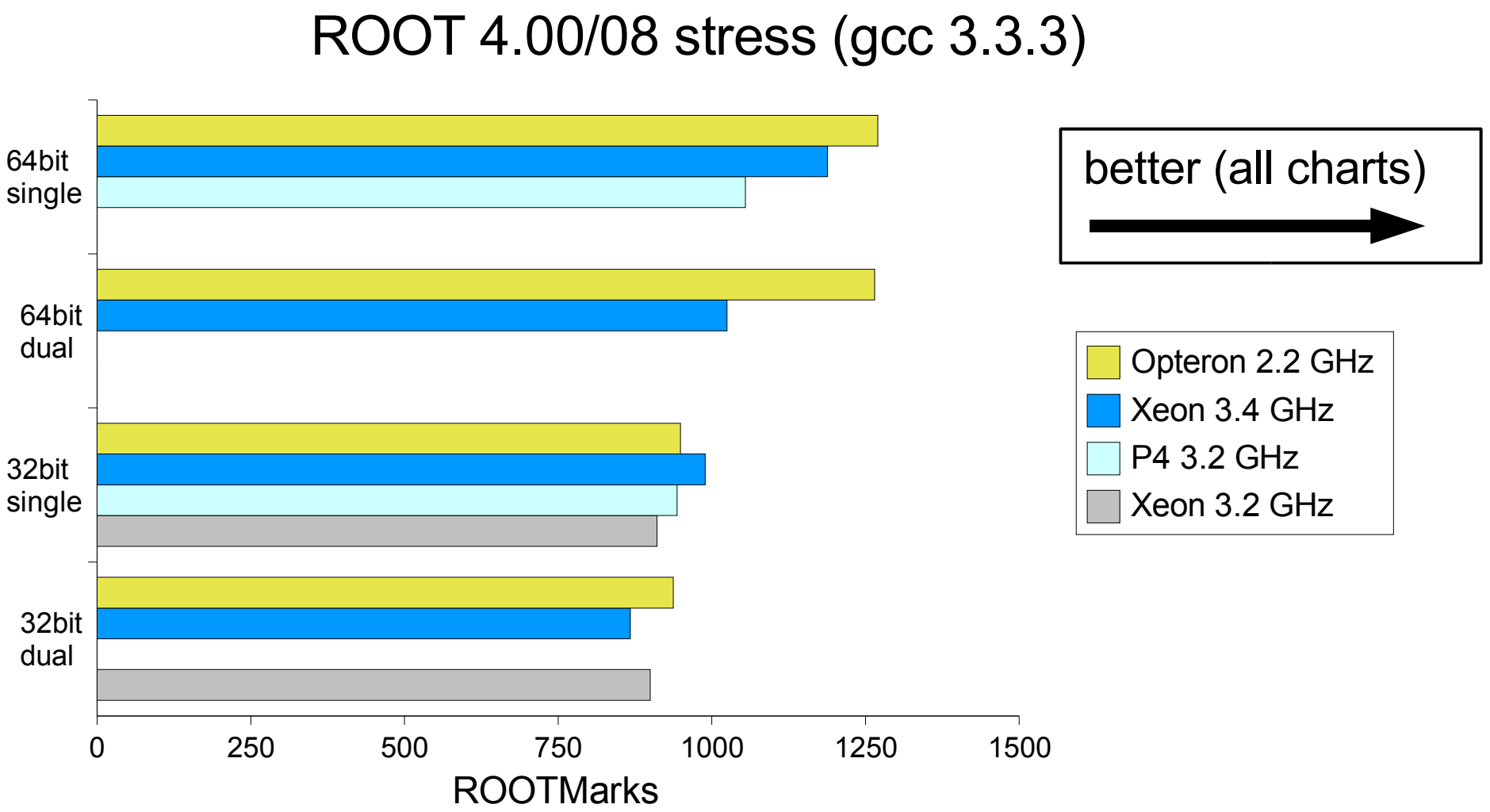
- All equipped with 2 CPUs and SCSI disk:
  - Opteron 2.0 GHz: IBM eServer 325, 4 GB
    - SuSE 9.0 professional, kernel 2.4.21-215-smp
  - Opteron 2.2 GHz: Sun Fire V20z, 4GB
    - SuSE 9.0 professional, kernel 2.4.21-231-smp
  - Xeon 3.4 GHz: Supermicro 7044H-X8R, 4GB
    - SuSE 9.1 professional, kernel 2.6.4-52-smp
  - Xeon 3.2 GHz: Sun Fire V65x, 2 GB, 32-bit
    - SuSE 8.2 professional, kernel 2.4.26
  - Tualatin 1.266 GHz: Supermicro 6013H, 1GB, 32-bit
    - SuSE 8.2 professional, kernel 2.4.25

EM64T arriving on the desktop

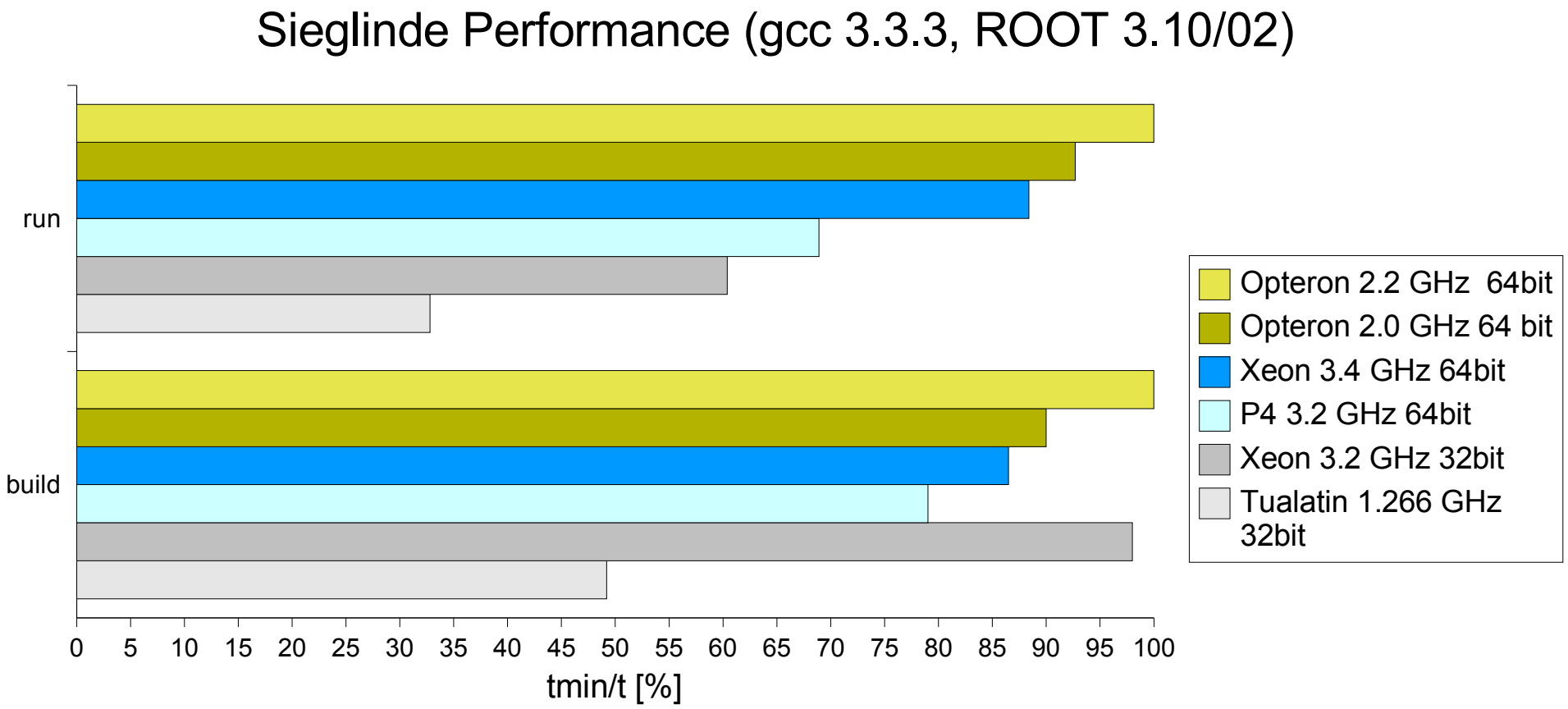
- your hands-on systems:
  - single P4 3.2 GHz
  - Dell Precision 370
  - 512 MB
  - SATA disk (80 GB WD)
  - SL 3.0.3, kernel 2.4.20-21.EL
  - 925X chipset

- yours are running 32-bit SL 3.0.4 though

ROOT Performance

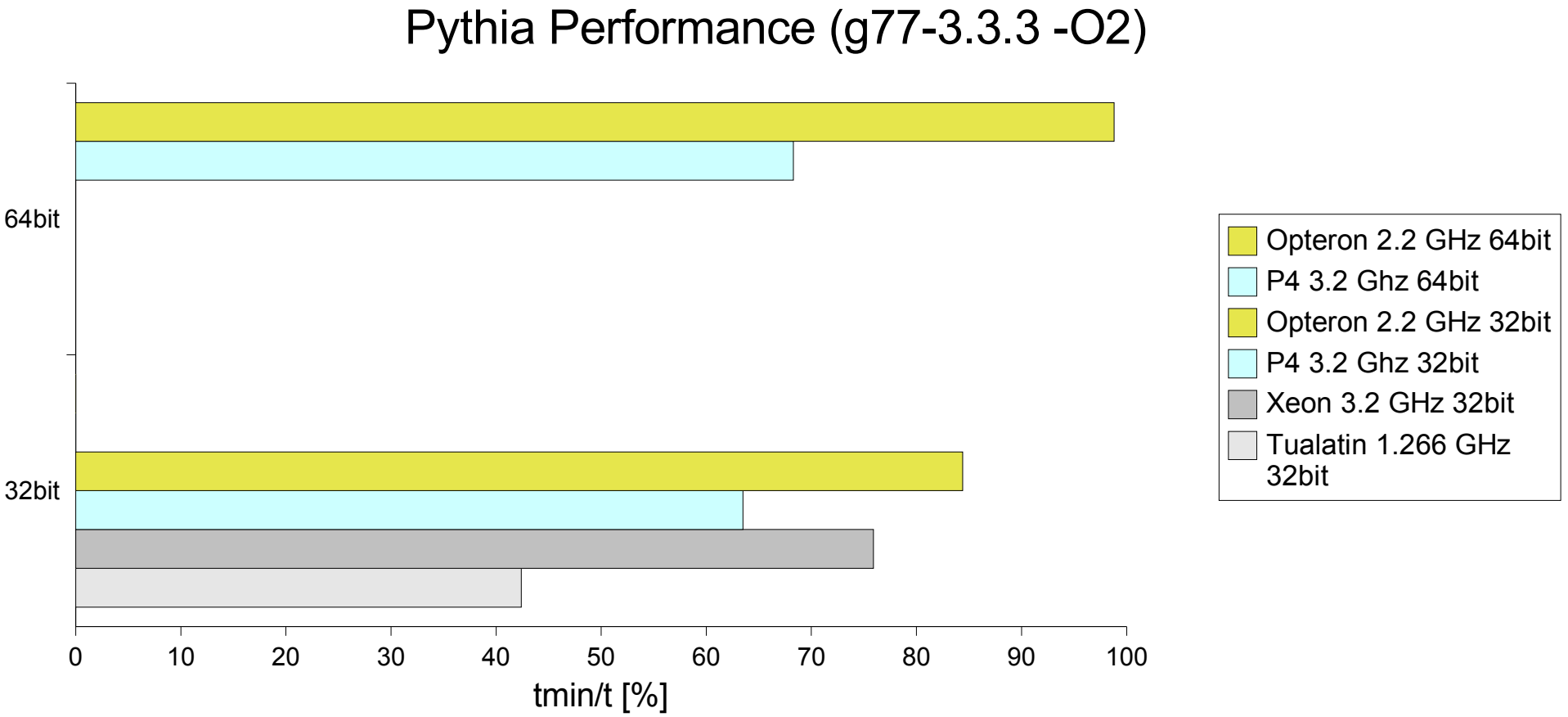


Sieglinde Benchmark



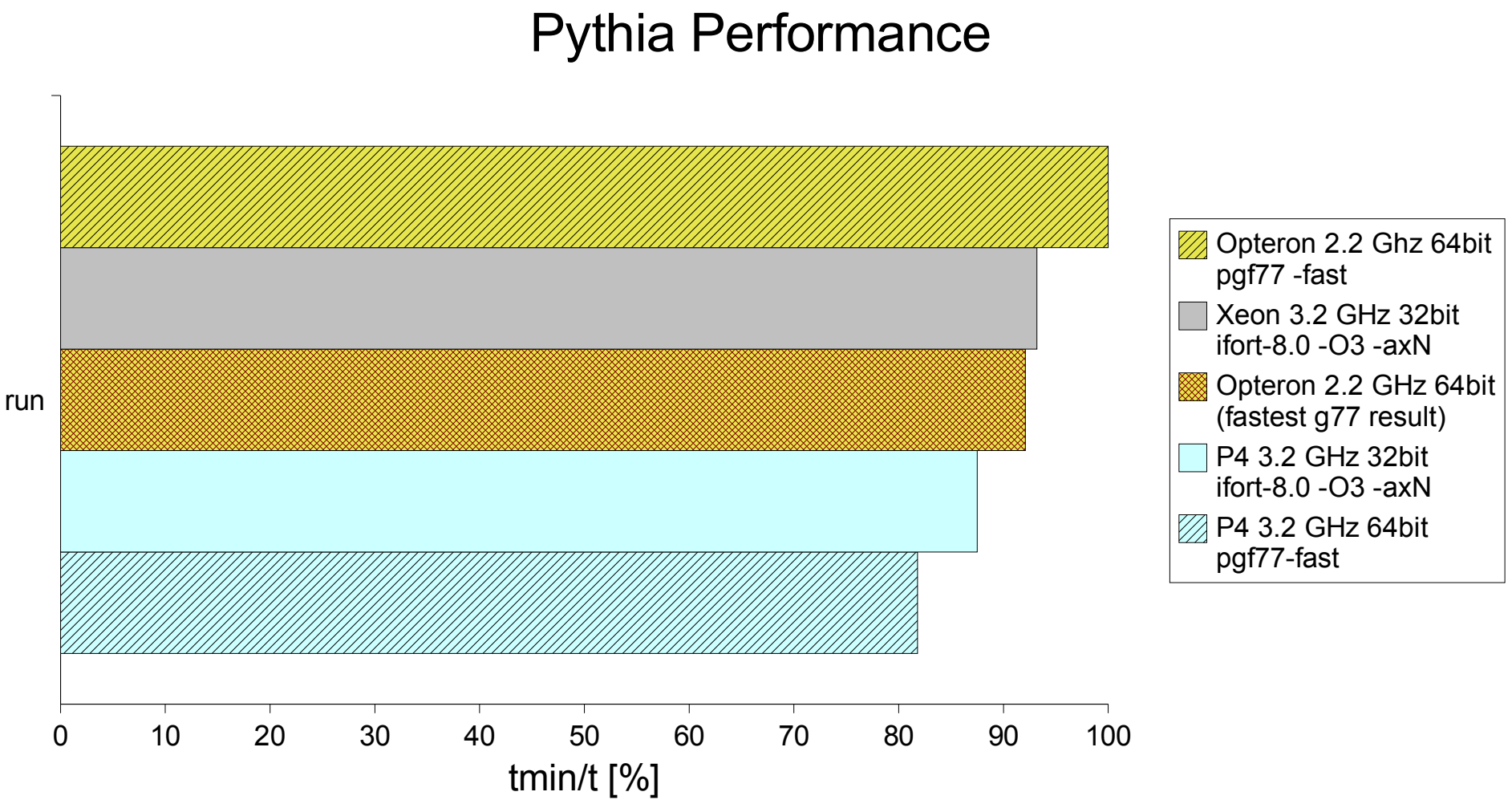
- Amanda experiment's neutrino reconstruction / filtering software
- single process, but uses a MySQL server on same host
- software made available by Peter Nießen, Univ. of Delaware

Pythia 6.2 (g77)

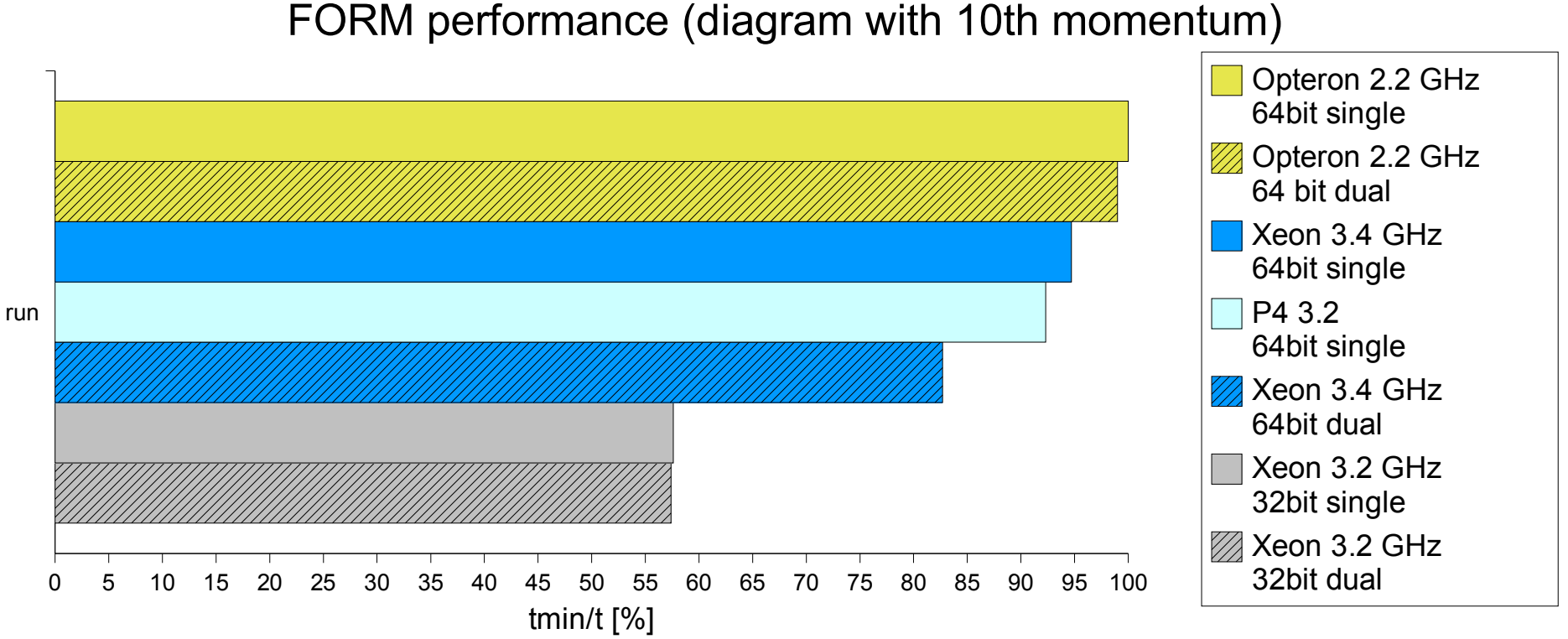


- Pythia 6.2 example 4
  "study of W mass shift by colour rearrangement at LEP 2"

Pythia 6.2 (Commercial Compilers)

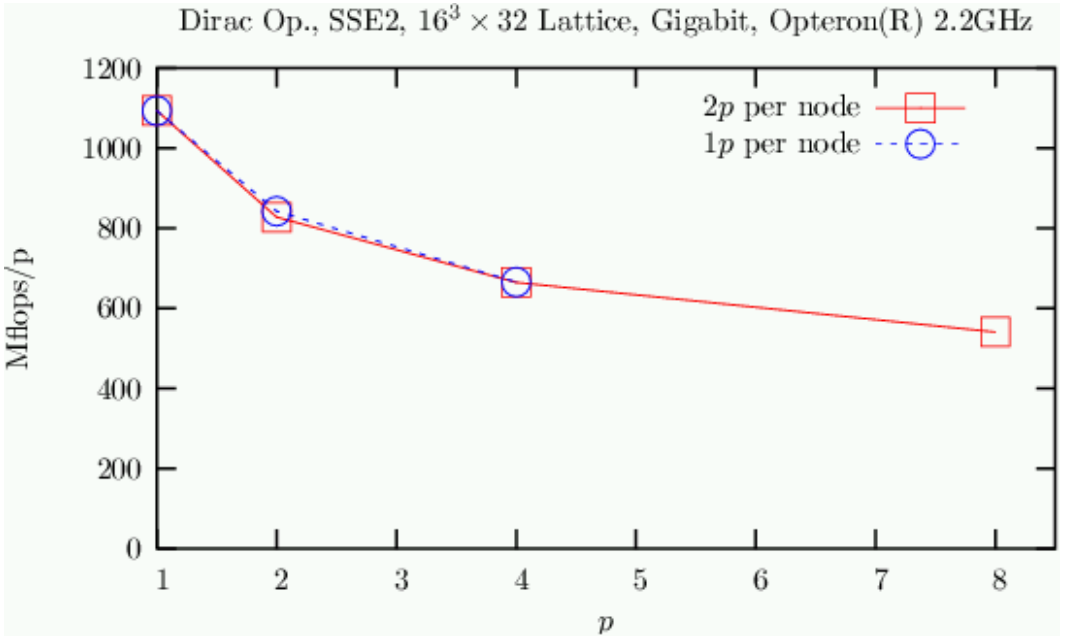


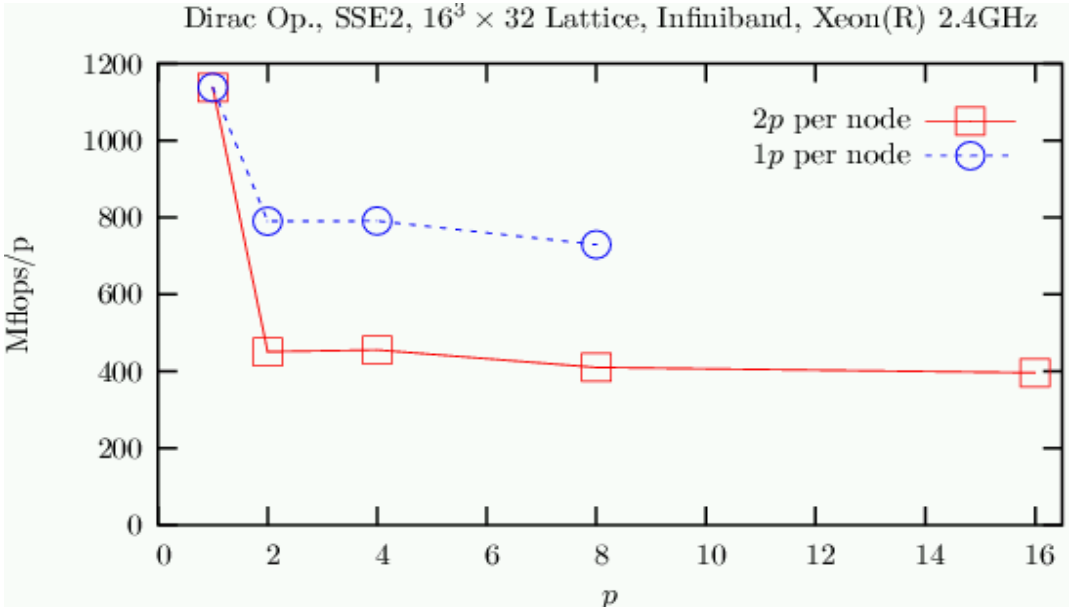
FORM 3.1



- symbolic formula manipulation, C, huge data sets
- implements own "paging" of data to disk
- 64bit executable built by author J. Vermaseren on DESY test system
- 32bit executable built with icc (www.nikhef.nl/~form)

Dual/Single CPU performance in Clusters



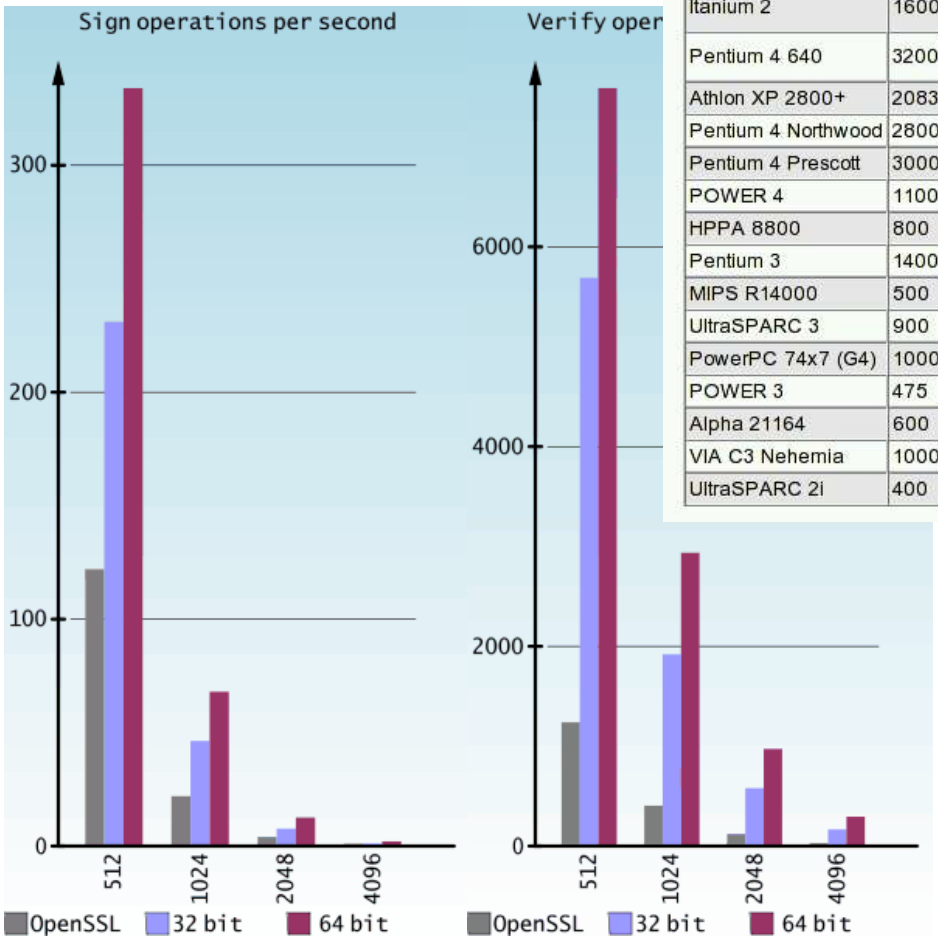


- measurements by C. Urbach, FU Berlin
- 32bit Lattice QCD, MPI
- performed on clusters with Gigabit Ethernet and Infiniband interconnects (FZK)
  - not our test systems
- p = number of processes

GNU MP

gmpbench 0.1 results
(www.swox.com/gmp/gmpbench.html)

CPU	freq MHz		Compiler/Compilation flags	base multiply	base divide	app rsa	GMPbench	Optimal (see note)
Opteron/Athlon64	2400	64	"gcc 3.4.2" -O2 -mcpu=nocona -funroll-loops **(NB! no asm code)**	27321	18280	1441	**5675**	11000
PowerPC 970 (G5)	2500	64	"gcc 3.4" -O3	20324	12874	1110	**4238**	7000
Alpha 21264	1000	64	"gcc 2.9-gnupro-99r1" -O2	16813	10706	782	**3240**	4500
Opteron/Athlon64	2400	32	"gcc 3.3.3" -O2 -fomit-frame-pointer **(NB! 32-bit only)**	19127	9823	802	**3316**	
Itanium 2	1600	64	"gcc 3.4.3" -O2 **(NB! Low-quality asm code)**	17046	9027	749	**3047**	10000
Pentium 4 640	3200	64	"gcc 3.4.2" -O2 -m64 -mtune=k8 **(NB! No asm code)**	14548	8973	779	**2984**	4000
Athlon XP 2800+	2083	32	"gcc 3.3.2" -O2 -fomit-frame-pointer	14076	7731	616	**2535**	
Pentium 4 Northwood	2800	32	"gcc 3.3.2" -O2 -fomit-frame-pointer -mcpu=pentium4 -march=pentium4	13013	5770	586	**2253**	3000
Pentium 4 Prescott	3000	32	"gcc 3.3.2" -O2 -fomit-frame-pointer -mcpu=pentium4 -march=pentium4	13348	5393	574	**2206**	3000
POWER 4	1100	64	"gcc 3.2.1" -O2 -maix64 -mpowerpc64 -mtune=power3	8951	5920	478	**1863**	
HPPA 8800	800	64	"cc B.11.11.30766" +DD64 +O2	9040	3724	362	**1450**	
Pentium 3	1400	32	"gcc 2.95.4" -O2 -fomit-frame-pointer	7097	3626	289	**1211**	
MIPS R14000	500	64	cc 7.3.0	5284	2819	241	**964**	
UltraSPARC 3	900	64	"cc 6.2" -fast -fns=no -fsimple=1 -xarch=v9 -xchip=ultra3	4242	2663	188	**796**	
PowerPC 74x7 (G4)	1000	32	"gcc 3.3.3" -O2 -mpowerpc	3453	2203	165	**676**	
POWER 3	475	64	"gcc 2.9-aix51-020209" -maix64 -mpowerpc64 -O2	3647	2259	157	**671**	
Alpha 21164	600	64	"gcc 3.2.1" -O2	3514	2185	158	**663**	
VIA C3 Nehemia	1000	32	"gcc 3.4.2" -O2 -fomit-frame-pointer -mtune=c3-2 -march=c3-2	2378	1314	111	**442**	
UltraSPARC 2i	400	64	"gcc 3.2.2" -O2 -mcpu=ultrasparc	1971	900	89	**343**	



- multiplication of large integers
- division by small primes
- RSA application (encryption)

Performance Comparisons: Summary

- AMD64/EM64T systems are fast, even in 32-bit mode
  - they're significantly faster in 64-bit mode
  - for the majority of codes - exceptions do exist
- Opteron systems make very efficient use of a 2nd CPU
  - and of additional MHz
- one gets more out of both with commercial compilers
- 64bit comes at a cost:
  - increased footprints in memory & on disk
    - typically 25%
  - different platform, and code may have to be fixed

64bit Linux on AMD64/EM64T systems

- good news: system looks, feels and behaves like "a linux PC"
  - BIOS (press F2 during boot...)
  - boot loader (grub, lilo)
  - OS installation (Red Hat, SL, SuSE)
- problems:
  - porting physics applications to 64-bit
  - providing/using 32-bit compatibility environments
  - residual bugs (features?) in 64-bit ports of system software

Porting Issues

- potential problems:
  - assumption that sizeof(int) = sizeof(long) = sizeof(void*)
  - inline assembly must not use x87 instructions
    - all FP calculations to be carried out in SSE registers
  - x87 registers were 80 bits wide
    - intermediate results kept in registers with this precision
      - was a problem when we moved from RISC to Linux/x86
    - intermediate results of FP arithemetics in SSE registers are 64bit again (standard IEEE 754 precision)
- can't mix 32/64-bit in same application
  - all libraries needed must be available as 64-bit

Now Shipping for AMD64/EM64T

- Oracle DB
- compilers, libraries:
  - Intel
  - PGI
  - NAG
  - Pathscale
- SUNs Java SDK 1.5 (5 ?)
- MySQL DB
- Mathematica 5, Matlab, ...
- Not yet: Maple

Data Type Sizes

type	x86	x86-64
char	8	8
short	16	16
int	32	32
long	32	64
long long	64	64
float	32	32
double	64	64
long double	96	128
void*	32	64

- no alignment constraints (like on i386)
  - but "natural" alignment is much faster

32bit Compatibility: Runtime

- 64bit linux allows running 32bit applications transparently
  - provided all shared libs are available
    - 64bit libraries go into .../lib64
    - 32bit libraries go into .../lib as before
    - mandated by Linux Standards Base, but not all ISVs comply
      - Oracle uses $ORACLE_HOME/lib and $ORACLE_HOME/lib32
    - some software (ROOT) needs completely separate install paths
  - some applications must be persuaded by using the "linux32" prefix command (see `setarch(1)`):
    - **`uname -m`** returns x86_64
    - **`linux32 uname -m`** returns i686
    - **`linux32 math`** (only app found to need this yet)

32bit Compatibility: Development

- 64-bit Linux also allows building 32-bit software
- gcc >= 3.2 creates 64bit objects by default on x86-64
  - **`-m32`** switch makes it create 32bit objects
    - and link against 32bit libraries
  - gcc3 on 32bit accepts the -m32 switch as well (noop there)
  - older versions need switches like `-Wa,--32` but still work
- reality is more complex
  - a decent Makefile uses commands like **`ginac-config --libs`**
- 32bit development best done in pure 32bit environment
  - may be chroot (or CHOS) environment on a 64-bit system

32bit Compatibility: Distributions

- Red Hat and SuSE (at least) provide 32bit packages
- the SuSE way:
  - RPM "xyz-32bit" with 32bit specific content
    - installed alongside the "xyz" 64bit package, no clashes
- the Red Hat way:
  - first install xyz.i386.rpm, then xyz.x86_64.rpm
  - (still limited) support by RPM/YUM, not yet by APT
    - **`rpm -ql glibc.i686`**
    - **`yum install openssl.i686; yum remove openssl.i686`**
  - reference counters for identical shared files
  - file/package color for clashing files (elf64 overrides elf32)

Technology Outlook

- the x86 architecture has not much headroom left
  - with or without 64-bit extensions
- CPU clocks won't become much faster
  - 4 GHz Xeons cancelled
- structures can't shrink much further
  - heat is the problem: approaching maximum density
- for years, you could rely on execution times going down by 20% every six months
- these times have passed

Overcoming the Limitations

- measures chip manufacturers are taking:
  - growing on-die caches (L2, L3):
    - 2 MB (now), 4 MB, ...
  - higher clock rates for buses and memory:
    - FSB1066, FSB1200,...
  - removing bottlenecks:
    - NUMA
- but most promising: multiple cores per CPU
  - dual core Opteron, Xeon (+HT!), Itanium this year
  - quad core Itanium in 2007
  - maybe even more

Multi Core CPUs: the Downside

- by the end of the year, a vanilla dual Xeon server will
  - have four cores
  - present eight virtual CPUs to the OS & application
- it will also require eight threads of execution for full utilization
- the same for a quad opteron with dual core CPUs
  - likely to be a common kind of system next year
- alas:
  - each core will be slower than today's CPUs
    - execution times will go up
      - unless problems can be parallelized - trivially, or otherwise

Summary & Conclusions

- 64-bit commodity hardware is a reality
  - AMD64/EM64T systems canonical choice for compute servers
    - rather sooner than later, 32-bit-only systems will vanish
- going 64-bit now will
  - give your application access to the top 20% of the potential
  - make the transition to IPF, once attractive, easy
    - young platform, moderate clocks, probably has some headroom
- other than those, do not expect a huge performance boost without effort, maybe for several years
- for more, you'll have to parallelize your applications
  - good price/performance soon for 8/16/32-way

Sources/Reading

- [1] Porting to AMD64 Frequently asked questions
  - http://www.amd.com/us-en/assets/content_type/DownloadableAssets/dwamd_AMD64_Porting_FAQ.pdf
- [2] The AMD64 ISA value proposition
  - http://www.amd.com/us-en/assets/content_type/DownloadableAssets/dwamd_Value_of_AMD64_White_Paper.pdf
- [3] Intel E7520/E7320 Product Brief
  - ftp://download.intel.com/design/chipsets/E7520_E7320/303033.pdf
- [4] Opteron 2P Server Comparison Reference
  - http://www.amd.com/us-en/assets/content_type/DownloadableAssets/30291C_brief_p1.pdf
- [5] Opteron 4P Server Comparison Reference
  - http://www.amd.com/us-en/assets/content_type/DownloadableAssets/30291C_brief_p1.pdf
- [6] Jan Hubička: Porting GCC to the AMD64 architecture
  - http://www.ucw.cz/~hubicka/papers/amd64.pdf
- [7] Bryan J. Smith: Dissecting PC Server Performance, SysAdmin Magazine November 2004
  - http://www.samag.com/documents/s=9408/sam0411b/0411b.htm